图书情报工作
第65卷 第9期 2021年5月

基于日志挖掘的学术搜索困难度量方法研究

■ 陈翀[1] 王思炜[1] 梁冰[2]

[1] 北京师范大学政府管理学院 北京 100875 [2] 中国科学技术信息研究所 北京 100038

摘 要：[目的/意义]用户检索中经常面临不同程度的信息搜寻困难，为更好地理解用户需求、改进检索系统，需要一种简洁有效的方法度量信息搜寻的困难程度。[方法/过程]将用户为查询而付出的行为及时间代价作为其信息搜寻困难的体现。按照用户在会话中的行为模式划分会话类型，将查询需求被满足且代价最小的会话类型作为比较基准，用基准会话的代价衡量其他会话类型的困难程度。为优化代价的表达模型，对搜寻代价的行为指标进行相关性检验，用因子分析选择独立性、区分度好的行为特征进行建模。以国家科技图书文献中心(NSTL)日志和搜狗日志为数据集比较学术搜索与通用搜索环境，以及不同会话类型所代表的探索过程中，用户的信息搜寻困难度。[结果/结论]在本文所度量的两种搜索系统中，用户面临的信息搜寻困难度分别为2.30和1.57，学术搜索中的困难高于通用搜索。在两种体现学术探索过程的会话中，困难度分别为2.35和4.13。本文提出的方法可以用简单的数值来概括具有多种影响因素的搜索困难，并能用于不同类型会话和搜索环境，丰富了检索系统的评估手段。

关键词：信息搜索困难 搜索困难度量 搜索代价 学术搜索 会话类型

分类号：G25

DOI：10.13266/j.issn.0252-3116.2021.09.009

1 引言

在获取信息时，人们会因为需求不明确、缺乏领域知识、或检索系统的性能欠佳、资源不足等各种原因而面临信息搜寻困难。在搜索过程中表现为用户频繁转换查询词、多次点击搜索结果、在结果详情的页面停留时间长等。学术搜索中的信息搜寻困难现象尤为常见。学术搜索是通过学术搜索引擎或文献数据库对学术研究领域的期刊、论文、作者等实体进行的检索[1]。在探索未知问题的过程中，学术用户需求不明确的情况非常普遍，但人们并不清楚用户到底面临多大的困难，也没有一个简单明了的方法度量这种困难。度量用户信息搜寻的困难是一种有用的系统评价手段，特别是对学术搜索这样的应用，有大量查询是源于不明确的需求，度量用户的困难可以帮助系统管理者或研究者更好地理解用户并改进系统功能。

学术搜索与通用搜索的差异主要体现在：①用户查询目的不同，前者为了科研探索，经常有较高比例的实体查询和对文献资源的下载需求[2]，而后者多用描述的方式表述需求，以浏览结果的方式获取信息；②查询词特点不同，学术搜索中的低频查询词更多，长尾分布更明显，且查询词经常是文献的题目、作者名称等，因而长度比通用搜索要长[2]；③点击浏览特点不同，通用搜索用户在搜索过程中倾向于点击浏览搜索结果列表中排名较后的结果[3]，而学术搜索用户会修改查询重新进行检索[4]。

本文认为，信息搜寻的难度体现为用户查找信息所花费的行为与时间代价，如查询、点击的次数，会话持续时间等。本文拟根据国家科技图书文献中心(National Science and Technology Digital Library，NSTL)和搜狗的用户交互行为日志，对完整搜索过程开展行为研究，并对比学术搜索与通用搜索环境下用户的行为特征，提出查询困难程度的衡量方式。需要说明的是，本文对学术搜索和通用搜索的比较及结论就是通过以上两个实际系统的对比得出。

2 相关研究

2.1 学术搜索和通用搜索中的用户行为特点

通过对学术搜索和通用搜索中用户行为研究的梳理，发现两者有较大差异。在查询特征方面，学术搜索

作者简介：陈翀(ORCID:0000-0002-9704-1575)，教授，博士，E-mail:chenchong@bnu.edu.cn；王思炜(ORCID:0000-0002-8646-9680)，硕士研究生；梁冰(ORCID:0000-0002-7622-6618)，高级工程师，博士。

收稿日期：2020-11-09 修回日期：2021-01-25 本文起止页码：79-88 本文责任编辑：易飞

的查询词长度、会话中的查询数、查询的持久性普遍大于通用搜索。学术搜索系统 ScienceDirect 和 CiteSeerX 中平均查询词长度为3.77[2]和4.76[5]个词语，而通用搜索系统 AOL 中的查询词长度为2.34[6]个词语，搜狗中为3.11[7]个词语。典型的通用搜索会话的查询次数小于2次[8]，但在学术搜索中，如 PubMed 用户会话的平均查询次数为4.44次[4]。R. I. Dogan 等[4]深入分析了 PubMed 中生物医学领域用户的行为特征，发现学术用户在信息搜寻中具有持久性，经常会进行查询重构，一般会按照作者名称、基因或蛋白和疾病名称等多种类型的实体进行搜索。在浏览下载行为方面，通用搜索中超过90%的用户只浏览前5个返回结果的详情[7]，而在学术搜索中，超过80%的浏览详情发生在前20个结果[4]。出于检索文献的需要，学术搜索用户最常见的操作依次是查询、下载、点击浏览，在有下载的会话中，平均下载次数是3次[1]。从行为表现上看，学术搜索用户在查询过程中更有耐心、并会付出更多代价。

以上研究使用的用户行为特征包括会话持续时间、查询频次、查询串长度、用户在结果列表中的点击位序、点击或下载的频次等，可为本文研究用户搜索困难选用合理的行为特征提供参考。用户与系统交互过程中的行为对用户检索体验有重要的指示作用。例如点击浏览或者下载可视为用户对检索结果满意的隐式反馈[9-10]。I. Yoo 等[11]认为理想的会话是查询一次就有浏览或下载行为，若经过几次查询后有浏览或下载行为，也可以视为用户信息需求得到满足的标志。

2.2 信息搜寻困难相关的用户行为研究

信息搜寻行为是用户在信息需求驱动下，在人机交互中表现出来的有目的性的系列活动[12-13]。搜索任务难度是指执行者对完成搜索任务所需工作量的评价[14]。难度可以通过客观度量来评估，比如任务执行时间和查询或浏览页面的数量，也可以通过主观度量来评估，比如用户对感知难度的自我评估[15]。

目前关于任务难度的研究主要集中在两个方面：一是任务难度对搜索行为的影响；二是根据用户行为对其是否面临困难进行判断。在难度对行为的影响研究上，J. Kim[16]在整个会话层面检验了任务难度与用户检索行为的相关性，发现不管是简单的事实性任务还是查询目标模糊的探索性任务，用户对困难程度的主观评价都与如下行为显著相关，包括：完成任务所需的时间、查询次数、浏览结果的数量。在搜索目标模糊、无明确边界的开放式任务中，用户经常进行查询重构，并花费较多的时间筛选检索结果[17]。总而言之，困难的任务与如下行为有相关性：更长的完成时间、更多的查询点击数、在搜索结果页面停留更长时间、甚至更多次放弃的查询[14,18]。以上结论为基于行为度量用户搜寻难度的研究提供了基础。在根据用户行为判断其面临搜寻困难的研究[14]上，本文与相关工作最大的不同是，后者主要是基于分类算法判断任务的难易[19]，不能对困难程度进行更细的区分。例如，A. Hassan 等[20]和 D. Odijk 等[21]通过分类算法区分了用户会话是属于艰难前进型，还是探索型。本文则是构建了一个困难程度计算模型，衡量搜寻困难程度，并且首次使用学术搜索和通用搜索两种日志数据度量用户在搜索过程中面临的困难。

关于搜索困难的研究方式主要有用户实验和日志挖掘两种。用户实验是在受控环境下观察用户完成不同难度搜索任务时的行为特点[22-23]，通过邀请被试完成事先设定的不同难度任务，分析用户在检索过程中自我评价的任务难度与其搜索行为之间的相关性。其优点是能够深入且有针对性地分析用户行为的影响因素，但实验条件与用户真实的检索情境还是存在差别。日志是分析用户搜索行为规律的重要依据，可以以会话为单位，研究用户在一次真实检索情境下的搜索行为特点。

2.3 日志中的会话划分研究

会话是用户为了特定信息需求与系统交互的过程，包括交互过程中产生的一系列行为[24-25]。会话划分对分析用户查询及其探索过程的完整性很重要。以往研究中，会话的划分主要有三种：最简单的是使用固定的持续时长[26]或相邻查询的时间间隔来划分，超过时间阈值的记录就被划分到下一个会话。时间阈值在不同的研究中使用不同的经验值，如通用搜索中的5分钟[8,27]、10－15分钟[28]等，OPAC日志中的30分钟[29]，购物日志中的45分钟[30]，等等。但按时间阈值的划分方式不一定适于学术搜索，因为学术用户的搜索兴趣可能很持久，按上述阈值分割日志会将对一个主题的查询行为分到不同会话片段，难以捕捉相对完整的搜索活动。为了使搜索过程被尽量完整地包括进会话中，有的研究以系统分配访问标识作为一个会话的标志，例如 T. Jiang 等[31]将用户开始对 OPAC 系统发出访问请求到离开系统之间的所有行为都归为一个会话，R. I. Dogan 等[4]则将用户在24小时内同一个 cookie ID 的行为记录都划入一个 PubMed 会话，并假设用户在一个会话中所有的查询都是与同一个主题相

陈翀，王思炜，梁冰. 基于日志挖掘的学术搜索困难度量方法研究[J]. 图书情报工作,2021,65(9):79–88.

关的。然而这种方式可能造成一个会话中混杂多个查询意图,于是有的研究又采用混合原则,例如 B. J. Jansen 等[24]采用 IP 地址、浏览器 cookie ID 加上相邻查询词相似作为会话划分依据,得到较好的分割效果,会话中相邻查询的主题一致性对会话划分有重要的影响。A. Hassan 等[20]考虑了时间间隔和查询主题一致性,将两个查询的时间间隔小于 10 分钟,且有共同的返回结果或相似的查询词的记录划分到同一个会话。

3 研究设计

3.1 基本概念和研究思路

本文关注学术搜索中用户信息搜寻困难的度量。一个会话 s 中包括用户为了特定信息需求与系统交互的一系列行为。学术搜索中的典型行为有:用户提交查询、在检索结果中点击浏览详情页面以及下载文献,分别记为 Q、B、D。用户的搜寻意图通过查询来表达,虽然查询的最终目标不一定是下载文献,但下载文献是最明确的判断依据。因此定义以下名词:

(1)达到目标。学术搜索中,用户的下载行为比查询和浏览行为更明确显示用户对检索结果的满意度[1,9-10],因此将会话中出现 D 作为用户达到搜索目标的一个行为标志。

(2)探索目标。是指用户逐步认识目标、明确查询需求的过程。典型表现是用户提交查询后,点击系统返回的结果列表并浏览某个结果详情,或调整表述再次查询。因此将会话中出现 Q、B 或只含 Q 作为探索目标的行为标志。

(3)搜寻代价。指为了探索或达到目标而在 Q、B 行为上花费的时间及频次开销。

(4)搜寻困难程度。本文通过用户在特定环境下付出的搜寻代价大小来反映。

研究思路为:首先进行会话划分,按照日志中同一用户的相邻查询串的相似性划分。其次定义会话类型,根据用户在探索目标和达到目标过程中的行为特征,将学术搜索会话划分为 S_0 到 S_3 4 种类型,如表 1 所示。其次,选取反映搜寻代价的行为指标,计算各类型会话中的行为特征,通过相关分析和因子分析,找出独立性好、具有代表性的基本特征对搜寻代价建模。再次,度量不同会话类型的搜寻困难程度,以搜寻代价最低且达到目标的会话类型为基准,其他类型会话的代价与基准代价相比,得到用户的信息搜寻困难度。最后,本文比较了学术搜索和通用搜索环境以及学术搜索的不同会话类型中用户所面临的困难,见图 1。

表 1 会话类型与含义

会话类型的含义	行为构成	会话中的行为序列	原子行为模式	经验认识
达到目标	S_0: {Q,D}	$q+\to d+(\to q\|d)^*$	$q\to d$	搜寻代价最低
探索目标	S_1: {Q}	$q+$	q	代价较低,直接放弃或浅探索
探索目标	S_2: {Q,B}	$q+\to b+(\to q\|b)^*$	$q\to b$	代价较高,深探索
探索并达到目标	S_3: {Q,B,D}	$q+\to d^*\to b^+(\to q\|b)^*\to d^+(\to q\|b\|d)^*$	$q\to b\to d$	搜寻代价较高

注:表 1"会话的行为序列"中"+"和"*"分别表示状态出现 1 到多次和 0 到多次

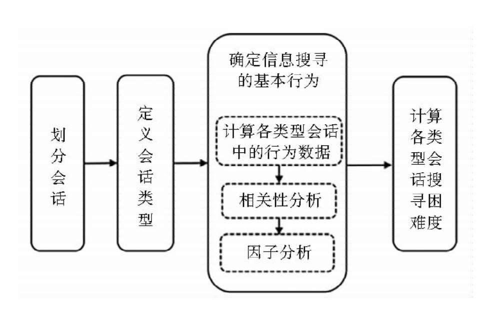


图 1 研究思路

3.2 会话的类型与搜寻代价

在通用搜索中,假设用户达到目标表现为找到特定的网页,理想情况是查询一次就找到目标,那么用户在达到搜索目标时搜寻代价最低的行为序列是 $Q\to B$,即"输入查询→浏览详情页面",而后结束任务;其他情况有:修改查询词重新检索、浏览,直到达到目标而结束任务或者直接放弃查询。

在学术搜索中,根据一个会话中包括的用户行为的不同,分为 4 种可能的会话类型:S_0、S_1、S_2、S_3。表 1 中是它们的含义及行为构成,用正则表达式归纳其行为序列,并给出最简单的原子行为模式及对其代价的经验认识。其中,根据 3.1 节对"达到目标"的定义,S_0 和 S_3 属于达到目标,S_1 和 S_2 是探索而未达到目标。前者按照搜寻代价分为两种,如果用户目标比较明确,例如要找一篇题目已知的文献,只经过 $q\to d$ 过程就可以达到目标,该类会话记为 S_0,搜寻代价最小;如果用户搜寻目标模糊或不能准确地表达信息需求,则会浏览详情甚至多次查询、浏览才会有下载,过程 $q\to b\to d$ 体现为探索,该类会话记为 S_3,搜寻代价较高。对未达到目标的两种会话 S_1 和 S_2,S_2 为 $q\to b$ 过程,表现为探

索；S_1 中用户查询后就结束了会话，行为中只有 Q，没有 B 和 D，即使用户对目标有所探索，程度也比较浅，付出的代价也相应较小。

定义单轮会话占比，可以用于解释用户对某一主题探索程度的“深”与“浅”。将仅包含原子行为模式的会话称为单轮会话，单轮会话说明其中的行为没有多次重复，用户没有在搜索中进行更多探索。在某类会话 S_i 中，单轮会话构成的集合记为 S_i'，$p_i' = |S_i'|/|S_i|$ 表示在第 i 类会话中，单轮会话数的占比。

3.3 信息搜寻的基本行为

会话中包含用户搜寻时的各类行为及其发生时间、行为的内容和结果等。表 2 将前人研究中采用过的会话行为归纳成查询 QF、浏览 BF、下载 DF、查询转换模式 QTF、时间 TF、以及点击结果位序 RF 等 6 种特征。其中下载特征 DF 中的 NumQueriesFirstDownload 和 NumBrowseFirstDownload 是指在第一次下载之前的查询数和浏览数，是反映用户探索初期确定目标的行为特征，也是本文借鉴相关工作，在学术搜索的下载行为特点中新增的指标。

上述指标并不相互独立，例如“会话 k 中所有查询的平均点击浏览位序”和“会话 k 中最大的点击浏览位序”有相关性。为衡量搜寻困难程度，对搜寻代价进行建模需要选取区分性和独立性较好的指标。为此，对表 2 中的指标进行因子分析。因子分析是利用降维的思想，从研究原始变量相关矩阵内部的依赖关系出发，提取数据中对变量起解释作用的少数公共因子。

表 2 所列查询特征 QF 中，CharQueryLen、WordQueryLen 是反映查询串的长度特征，查询转换特征 QTF 中，AvgQuerySim、AddTerms、DelTerms 是考察相邻查询串相似度的指标，它们体现的是查询内容的变化，不直接反映搜寻行为代价，因此本文选取其余 10 个行为指标进行相关分析和因子分析。最终找出与其他行为相关度较高的公共因子，根据这些因子具有的含义选取对应的行为指标，构成基本行为集合。

表 2 会话中的搜索行为特征指标

Type	Name	Description
Query Features(QF)	NumQueries[14,20-21]	Number of queries in a session
	AvgCharQueryLen[19-20]	Average number of characters of queries in a session
	AvgWordQueryLen[20-21]	Average number of words of queries in a session
Browse Features(BF)	NumBrowse[20-21]	Number of browse in a session
	BrowsePerQuery[20]	Average number of browse per query
	NumQueriesFirstBrowse[19]	Number of queries before the first browse
Query Transition Features(QTF)	AvgQuerySim[20]	Similarity between adjacent queries
	AddTerms[20-21]	Number of added terms
	DelTerms[20-21]	Number of removed terms
Time Features(TF)	Total DwellTime[1,20]	Total dwell time in a session
Rank Features(RF)	MaxRanks[7,21]	Maximum ranks of the browsed results of queries in a session
	AvgRanks[7,21]	Average ranks of the browsed results of queries in a session
Download features(DF)	NumDownloads[1]	Number of downloads in a session
	NumQueriesFirstDownload	Number of queries before the first download
	NumBrowseFirstDownload	Number of browsed pages before the first download

3.4 信息搜寻困难的度量模型

基于用户日志研究搜寻困难的前人工作并不多。A. Hassan 等[20]和 D. Odijk 等[21]通过日志中的用户行为特征来判断搜索会话是属于探索还是陷入困难，但并未对搜索困难的程度进行度量。本文根据基本行为的代价对信息搜寻的困难程度建模，可对不同会话、不同会话类型以及不同系统的搜索代价进行比较。考虑到学术用户探寻式搜索较多，为了对学术搜索中用户面临的困难有一般性的认识，本文首先从会话类型层面比较度量，进而是系统层面。

（1）比较基准。选择用户达到目标时付出搜索代价最低的会话类型，即 S_0。S_0 中没有反复浏览详情筛选结果的行为开销。

（2）基本行为集合。以反映搜索代价，并且具有相互独立性和代表性的行为指标构成的集合，记为 T，第 j 种基本行为为 $t^j \in T, j \in [1, M], M = |T|$。

（3）会话类型层面的度量模型。设特定应用中的会话类型构成集合 S，$S = \{S_0, \cdots, S_i, \cdots\}, i \in [1, N]$，$N$ 是除 S_0 之外的会话类型数。第 i 类会话 $S_i \in S$ 且 $S_i = \{s_{i1}, \cdots, s_{ik}, \cdots, s_{in}\}$，$s_{ik}$ 代表 S_i 中第 k 个会话。设 t_i^j 代

表第 j 种基本行为在 S_i 类会话上的平均代价。

$t_i^j = 1/n\ \sum_{k=0}^{n} t_{i,k}^j$

代价以具体行为的时长或频次表示。S_i 类会话中用户面临的信息搜寻困难程度可以用搜索代价 c_i 计算，即通过基本行为的代价 t_i^j 与 S_0 中对应的行为代价 t_0^j 之比加权得到，其计算公式为：

$$c_i = \frac{1}{M}\sum_{j=1}^{M}\frac{t_i^j}{t_0^j} \qquad 式(1)$$

（4）系统层面的度量模型。由 S 中每种会话类型的数量占比及其对应的困难程度加权组成，其计算公式为：

$$F = \sum_{i=0}^{N} p_i * c_i \qquad 式(2)$$

其中，p_i 是 S_i 类会话在整个系统中的数量占比。需要说明的是，如果比较基准选择的是代价最小的单个会话 s_0，也可以按上述思路计算单个会话 s_j 的困难程度，此时 t^j 就是单个会话中的基本行为的数值而非平均值。

4 实验

4.1 实验数据及预处理

本文分别选取 NSTL 和搜狗代表学术搜索系统和通用搜索系统，对前者日志选择 2017 年 7 月 1 日至 28 日间的检索、浏览和下载记录，约 22.6 万条，各字段及含义见表 3；后者日志为搜狗实验室 公开发布的搜狗搜索引擎 2008 年 6 月某 24 小时内的访问数据，约 170 万条记录，各字段及含义见表 4。其中，搜狗日志中的 ID 标记了用户从开始访问到离开系统之前行为的集合，虽不对应单个查询主题，但能够标识特定用户的一次访问。而 NSTL 中的 UserID 有所不同，科研机构会购买集团账号供内部用户使用，因此 UserID 不能指代特定个体，即使日志中有同一个 UserID 的连续查询，也有可能是多个用户发出的，其查询内容自然不相同。为此本文借鉴 B. J. Jansen 研究中的做法[24]，采用 UserID + IP 区分用户，对同一个 UserID 且相同 IP 的记录，人工判断查询的主题一致性，如果相邻查询至少共享一个完全相同或者相近的非停用词，则被划分到同一个会话，否则分到两个相邻会话。对 NSTL 日志进行人工标注后得到会话 2 933 个，包括 15 372 条行为记录。对搜狗日志，剔除重复的 ID 记录和异常查询后，得到 63 060 个唯一 ID，随机采样抽取其中 5% 的 ID，人工标注得到 8 324 个会话，包括 36 752 条行为记录。会话类型见 3.2 节表 1。

通用搜索用户找到目标后一般是点击结果浏览而非下载，因此日志只包含 Q、B 两种行为。搜狗数据又过滤了仅含查询的条目，所以实验所用的通用搜索日志的会话行为模式比较单一，但又不同于学术搜索中 S_2 类型，因为通用搜索中的 $Q \rightarrow B$ 更可能代表查询成功。为了与学术搜索的 S_2 区分，将通用搜索会话标为 S_4 类型。

表 3　学术搜索日志主要字段名称与格式示例

字段名字	字段含义	格式示例
UserID	用户名	字符串
IP	IP 地址	*.*.*.*
QueryDate	查询时间	2017/7/20 15:03:24
Type	访问文献的类型	[J02]//表示 English Journal
QueryWords	查询表达式	字符串，如"Big Data"
BrowseDate	浏览时间	2017/7/20 15:05:24
DocID	文章编号	32 位 16 进制
URL	访问 URL 信息	如 nstl.gov.cn/search.html? t=…
RequestDate	请求下载时间	2017/7/20 15:13:12

表 4　搜狗搜索日志主要字段名称与格式示例

字段名称	字段含义	格式示例
QueryDate	用户查询或点击的时间	01:22:34
ID	由系统自动分配的 session 标识号	18630035486060958
QueryWords	用户提交的查询式	URL 或字符串
Rank	该 URL 在返回结果中的排名	1
ClickRank	用户的点击结果中的顺序号	1
URL	用户点击的 URL	www.***com/index.php

4.2 学术搜索和通用搜索中的用户行为

（1）不同会话类型的行为特征。表 5 是学术搜索各类会话的数量分布 p_i 以及单轮会话在同类型会话中的占比 p_i'。表 6 给出了学术与通用搜索中不同类型会话对应的搜索行为特征平均值，其中的指标来自表 2。

表 5　学术搜索中的会话类型及单轮会话占比

会话类型	$\|S_i\|$	p_i/%	p_i'/%
S_0:{Q,D}	538	18.34	84.20
S_1:{Q}	1 314	44.80	76.48
S_2:{Q,B}	717	24.45	24.97
S_3:{Q,B,D}	364	12.41	37.09

基准会话 S_0 代表在学术搜索中能直接达成查询目标的会话，但根据表 5，S_0 型会话只占 18.34%，其中单轮会话占 84.20%。也即 NSTL 系统数据中只有 15.44% 的学术搜索查询需求明确且直接命中资源。根据表 6，S_0 型会话的平均查询长度为 83.55 个字符，比 S_1、S_2、S_3 都高，这说明需求明确的学术用户是直接以文献标题搜索并下载文献。而且学术搜索各类型会话的平均查询词长度都远高于通用搜索，这是因为前

表 6 不同类型会话的搜索行为特征平均值

Type	Name	S_0	S_1	S_2	S_3	S_4
Query Features(QF)	NumQueries	1.36	1.52	3.36	4.34	1.80
	CharQueryLen	83.55	55.46	23.66	48.46	8.40
	WordQueryLen	8.41	6.27	3.74	5.82	1.89
Browse Features(BF)	NumBrowse	——	——	4.15	9.44	4.41
	BrowsePerQuery	——	——	1.68	3.18	2.27
	NumQueriesFirsBrowse	——	——	1.70	1.57	——
Rank Features(RF)	AvgRanks	17.63	18.23 *	25.00	24.45	8.31
	MaxRanks	20.15	20.77 *	38.64	41.04	13.34
Download Features(DF)	NumDownloads	1.84	——	——	3.37	——
	NumQueriesFirstDownload	1.22	——	——	2.98	——
	NumBrowseFirstDownload	——	——	——	5.51	——
Time Feature(TF)	Total DwellTime(one day)/分钟	20.80	7.50	77.60	90.50	38.20
	Total DwellTime/分钟	525.60	178.60	721.70	1 307.60	38.20

注:由于搜狗日志只有 24 小时内的数据,为了便于比较学术搜索和通用搜索的用户行为,Total AvgDwellTime(one day)是只考虑在 24 小时内结束的会话的平均时长。表中 * 号标的 S_1 型会话的点击位序特征 *RF* 体现了用户在返回结果中有翻页,但这不同于其他会话类型中的点击位序,S_1 中没有行为 *B*。在其他类型会话中,*RF* 特征的值代表用户为了浏览详情而点击的结果位序

者用户以找文献资源为主要需求,用户更愿意提交较长的查询字符串。

S_1 型会话只包含查询行为,无浏览和下载,其中单轮会话的占比为 76.48%,说明大部分用户只提交一次查询就结束会话。只有少数会话对同一主题进行了多次查询,会话中的平均查询次数为 1.52。这类会话中的探索行为较少,本文不考虑这种会话的代价。

S_2 和 S_3 型会话体现探索过程,据表 5,它们在会话中占 24.45% 和 12.41%,但单轮会话的占比只有 24.97% 和 37.09%,表 6 显示这些会话的平均查询次数为 3.36 和 4.34 次,浏览的结果在检索结果列表的平均位序为 25 和 24.45。特别是 S_3 型会话中,NumBrowse 为 9.44,远高于通用搜索的 4.41。这说明学术搜索用户探索过程相对深入,S_3 尤其典型,在查询未能直达目标的情况下,用户仍愿继续付出搜寻代价。相比而言,通用搜索的 S_4 型会话中平均查询只有 1.8 次,平均点击浏览的位序为 8.31。

(2)会话持续时长。会话持续时长是用户探索深入程度的重要指标。表 6 中 Total DwellTime 统计了各种类型会话的平均持续时长,顺序为 $S_3 > S_2 > S_0 > S_1 > S_4$。由于存在一些持续时间超过几天的异常会话,本文采用两种处理方法,一是取在 24 小时内结束的会话,平均持续时长 Total DwellTime (one day)为 $S_3 > S_2 > S_4 > S_0 > S_1$,具体可见表 6;二是按持续时长排序后,取数量为前 80% 的会话,仍得到上述次序,且 80% 的 S_0 型会话持续时间不超过 6.92 分钟,S_1 型为 0.47 分钟,S_2 型为 107.87 分钟,S_3 型为 589.08 分钟。S_4 为 20.60 分钟,如图 2 所示:

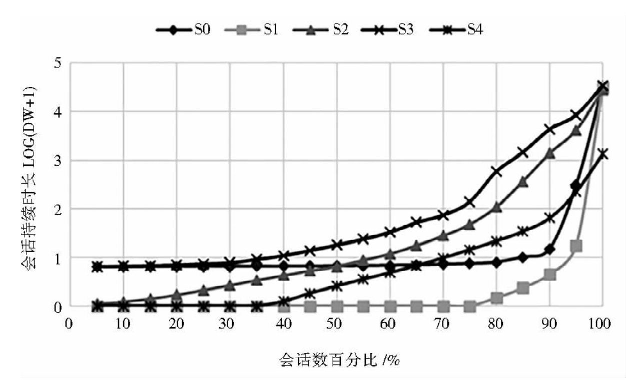


图 2 各类型会话持续时间分布

图 2 是各类型会话持续时间分布。会话按持续时长 DW 升序排列,由于差异较大,为了便于作图,用 y - 对数坐标系。横坐标为会话数百分比,纵坐标为 log(*DW* + 1)的值。如果只考虑前 80% 的会话,可以看出 S_0 型会话的时长短且数值稳定。S_1 时长极短,且曲线低缓,说明这类型会话的用户未对查询主题进行深入的探索。S_2、S_3 和 S_4 的会话时长都随会话数量百分比上升而递增明显,但前两条曲线在 S_4 的上方,这体现了学术搜索用户在探索主题时的耐心和查询持续性高于通用搜索,另一方面从 S_2、S_3 的耗时也说明学术搜索的代价高、困难程度较大。

4.3 基本行为特征的选择

对表 6 中所列行为特征进行相关分析和因子分析,选择反映搜索代价的基本行为特征。

4.3.1 相关性分析

由于 S_3 的探索特点明显,数据较全,本文选用 S_3 会话研究。对每组表示行为特征的数据进行单样本的 Kolmogorov-Smirnov 检验,由于各组数据的显著性均小于0.05,不服从正态分布,因此采用 Spearman 检验各组数据之间的相关性,结果见表7,各行为特征之间的相关性大于0.5 的值在下三角区加粗标出。

表 7　10 个行为指标的相关性分析结果

characteristic variables	相关系数									
	NumQueries	NumBrowse	NumDownloads	BrowsePerQuery	Total DwellTime	NumQueries FirstDownload	NumBrowse FirstDownload	NumQueries FirsBrowse	AvgRanks	MaxRanks
NumQueries	1	.623 **	.565 **	－.495 **	.662 **	.858 **	.483 **	.547 **	.612 **	.664 **
NumBrowse	**.623** **	1	.618 **	.213 **	.564 **	.526 **	.875 **	.163 **	.564 **	.577 **
NumDownloads	**.565** **	**.618** **	1	0.097	.480 **	.471 **	.487 **	.179 **	.617 **	.623 **
BrowsePerQuery	－.495 **	.213 **	0.097	1	－.210 **	－.421 **	.287 **	－.2 **	－0.049	－.110 *
Total DwellTime	**.662** **	**.564** **	.480 **	－.210 **	1	.492 **	.427 **	.320 **	.456 **	.479 **
NumQueriesFirstDownload	**.858** **	**.526** **	.471 **	－.421 **	.492 **	1	.492 **	.659 **	.592 **	.634 **
NumBrowseFirstDownload	.483 **	**.875** **	.487 **	.287 **	.427 **	.492 **	1	0.056	.520 **	.524 **
NumQueriesFirsBrowse	**.547** **	.163 **	.179 **	－.402 **	.320 **	**.659** **	0.056	1	.293 **	.328 **
AvgRanks	**.612** **	**.564** **	**.617** **	－0.049	.456 **	**.592** **	**.520** **	.293 **	1	.989 **
MaxRanks	**.664** **	**.577** **	**.623** **	－.110 *	.479 **	**.634** **	**.524** **	.328 **	**.989** **	1

注:**表示在置信度(双测)为 0.01 时,相关性显著;*表示在置信度(双测)为 0.05 时,相关性显著

4.3.2 因子分析

选取包含行为种类最多的 S_3 型会话,对表7中的10种行为的数据进行因子分析,按照特征值大于1的标准提取公因子,得到3个公因子。在 KMO 和 Bartlett 的检验中,KMO 的值为 0.600,Bartlett 球形统计量为 1 841.620,Sig. 的值小于 0.05,因此,样本数据适合做因子分析。当公因子为3时,共解释了 68.08% 的方差,公因子的数量为4时,共解释了 77.52% 的方差。表8是旋转成分矩阵,各列是4个公因子。10个行为变量在各个因子中的系数大小表示变量在因子中的载荷大小,其绝对值越大,说明该变量与公因子的相关程度越高,该变量应该归属于这个公因子,因此,本文选择每个公因子中载荷系数较大的变量来代表该公因子,最终得到计算信息搜索代价的基本行为集合 T＝{NumBrowse、NumQueries、AvgRanks、Total DwellTime}。

需要说明的是,NumQueries 和 NumQueriesFirstDownload 这两个变量对公因子2的影响非常接近且均较大,考虑到并非所有的会话类型都有下载行为,本文中只有 S_0 和 S_3 型会话才有 NumQueriesFirstDownload 的值,为了兼顾计算需要,选择另一个变量 NumQueries 作为基本行为。

表 8　旋转成份矩阵

基本行为变量	成份(公因子)			
	1	2	3	4
NumQueries	.141	**.850**	.161	.280
NumBrowse	**.925**	.067	.023	.050
NumDownloads	.474	.229	.513	.078
BrowsePerQuery	.734	－.249	.460	－.051
Total DwellTime	.006	.101	－.008	**.955**
NumQueriesFirstDownload	.013	**.888**	.125	.162
NumBrowseFirstDownload	.784	.201	.005	－.025
NumQueriesFirsBrowse	.069	.712	－.051	－.178
AvgRanks	.084	－.090	**.922**	－.050
MaxRanks	.043	.241	.892	.041

4.4 信息搜寻困难程度

4.4.1 各种类型会话的信息搜寻困难度计算

由于日志数据本身的局限,部分数据需要进行缺失值的合理填充和平滑。S_0 中没有浏览行为 B,用户提交查询 Q 后,直接产生下载行为 D,假设用户对下载结果的判断相当于一次页面浏览,则可以把表9中 S_0 会话的 NumBrowse 平均值设为1。另外,NSTL 日志中记录的是用户对查询结果的翻页数,并未记录用户所浏览页面的具体位序,本文按照 NSTL 系统每页默认的显示结果数10条/页与翻页数进行平滑。如果一个 NSTL 会话中有3次查询行为 Q,分别为 q_1、q_2、和 q_3,各查询的翻页数分别为1、2、3,则会话中3个查询对应的点击位序分别算作15、25、35,则得到该会话的 AvgRank 为25,MaxRank 为35。按照3.4节公式1和4.2节表6,求得各类型会话的基本行为均值及会话代价数据,见表9。

4.4.2 学术搜索与通用搜索中的信息搜寻困难程度

按照公式(2)计算学术搜索平台中用户信息搜寻

表 9 各类型会话的基本行为均值及会话代价

公因子	行为指标	S_0	S_2	S_3	S_4
H_1	NumBrowse	1.00	4.15	9.44	4.41
H_2	NumQueries	1.36	3.36	4.34	1.80
H_3	AvgRanks	17.63	25.00	24.45	8.31
H_4	Total DwellTime	525.60	721.70	1307.60	38.20
会话代价		1	2.35	4.13	1.57

的平均困难程度 F。选取 S_0、S_2 和 S_3 型会话，参照表 5 中的数值，并从总量中刨除 S_1 重新计算占比，即得到 $p_0=0.332$，$p_2=0.443$，$p_3=0.225$，则得到本文所观察的学术搜索平台 NSTL 中用户的信息搜寻困难度为 $F_A=2.30$。如果将本文中使用的通用搜索平台日志会话视为只有 S_4 这一种情况，则得到通用搜索平台的搜索困难度为 $F_G=c_4=1.57$。

5 分析与讨论

5.1 学术搜索的会话划分

从图 2 可知，只有给出查询并下载文献的类型 S_0 和放弃查询的会话类型 S_1 具有较稳定的时长。80% 的 S_1 型会话持续半分钟以内，80% 的 S_0 型会话持续 5－7 分钟。而其他类型的会话出于对查询主题的探索需要，都具有较大的时间差异。因此如果对用户一个查询过程进行研究，按照时间间隔划分会话的做法不可取。无论是通用搜索还是学术搜索，都很难找到一个合理的经验值。按照查询主题进行会话划分更合理。本文由人工按照查询主题的一致性划分，以保证对特定主题的查询和探索行为不被割裂。

5.2 学术搜索中的用户特点

总体上讲，学术搜索用户对所查询主题的耐心比通用搜索用户更大，愿意付出更大的代价，行为上表现得比较复杂。本文按行为模式不同将学术搜索概括为 4 种类型。反复查询和浏览详情可视为典型的探索行为，在 S_2、S_3 型会话中比较普遍，从单轮会话占比 24.97% 和 37.09% 可看出，多数情况下用户在通过反复查询、浏览和下载来筛选判断。此外，学术搜索中的会话持续时间普遍较长，S_2 型会话虽未找到下载目标，但也有将近 30% 的会话持续时长大于 30 分钟。在通过查询和浏览详情达到目标的 S_3 型会话中，将近 50% 的会话时长大于 30 分钟。而通用搜索环境中，80% 的会话在 20 分钟以内结束。

5.3 S_1 型会话放弃查询的可能原因

S_1 型会话中，用户停止继续搜寻的原因或许是查询目标仅为寻找文献标题，需求得到了满足，但更可能是用户放弃了查询。放弃查询与系统的资源量、性能、用户自身的知识等都有关系。从表 6 看，S_1 型会话的查询词平均长度 WordQueryLen 为 6.27 个词语，与 S_0 会话的 8.41 个词语相差不多。S_0 会话的查询有需求明确、描述详细的特点，S_1 型会话的查询也可视同。但 S_1 型会话中的平均查询次数 NumQueries 仅为 1.52 次，用户并未尝试多次查询，说明用户放弃的原因可能是系统中不包括用户所需资源，导致返回结果与目标差别很大，用户判断不必翻页、浏览或再次查询，就转而放弃。由于本文所用数据中 S_1 型会话占比较大，达到 44.80%，侧面说明了 NSTL 系统的资源量有待扩充。

5.4 学术搜索系统改进分析

S_2 和 S_3 型会话较能体现用户持续探索的特点，它们的总数占系统会话数的 36.86%，多数会话不是单轮的，检索过程经历了曲折。用户反复探索固然有自身需求复杂的原因，但从 S_2 型会话的 WordQueryLen 为 3.74 个词语来看，这类会话的查询对需求的描述不足，系统可以用查询词推荐功能辅助用户表述需求，减少用户描述需求的模糊性。此外，达成目标的 S_0、S_3 型只占总数的 30.75%，S_3 型会话的平均查询数、浏览数和下载数分别为 4.34、9.44、3.37，大致上每个查询对应 3.18 次浏览详情，产生 1 次下载，会话中点击结果的最大位序平均是 41.04。为了减少用户浏览详情的时间开销，系统可以通过用户点击结果改进排序算法，降低搜寻信息的困难程度。

6 总结与展望

本文探讨了学术搜索和通用搜索中用户的信息搜寻困难。首先，根据学术搜索中用户的行为构成划分会话类型，分析每种类型会话中用户的搜索代价。其次，通过相关性分析和因子分析找出能反映搜寻代价的基本行为特征，构建模型来度量不同会话类型以及不同搜索平台中用户的信息搜寻困难程度。以学术搜索中达到搜索目标且搜索代价最小的会话类型 S_0 为基准，得到学术搜索中用户的搜索困难程度为 2.30；通用搜索与之相比，用户的搜索困难程度为 1.57。可以得出用户在学术搜索中面临的平均困难程度大于通用搜索。

以往关于用户搜索困难的研究多是基于用户行为实验，而用户行为实验是在受控的情况下进行的，研究人员需要提前设定用户的检索任务。而在实际检索中用户的搜索任务是潜在的、无法预先观察的，并且可能存在一段时间内交替进行多个检索任务的情况。在用

户实验中,用户检索的困难程度是通过检索任务本身难度和用户自我评级确定,存在主观标准不一致的情况。因此本文可以为后续基于日志数据研究用户的搜索困难提供新的思路。

本文的主要贡献有:①提出一种信息搜索困难程度的度量方法,可用于评价不同搜索环境下用户的信息搜寻困难。丰富了评价指标,当系统有功能改进时,能较为容易地通过计算改进前后用户搜索会话的困难程度来评价系统的改进效果。②针对用户行为复杂的学术搜索环境,提出根据行为模式划分会话类型的方法,有助于更细致地比较不同搜寻情况下用户面临的困难,可以依据会话类型有针对性地改进功能。③根据对学术搜索与通用搜索中用户信息搜寻行为代价的比较,得出学术搜索中用户对查询主题的兴趣更持久,搜索过程显得更具耐心等认识。④本文中的会话是人工按主题一致性标注的,有较好的准确度和可靠性,能为会话识别研究提供评测数据。

本文还存在一些不足,在根据用户的行为特征确定用户会话是否达到目标时,因为不能根据日志数据获取用户真实的查询意图,所以以下载作为达到目标的标志。这个假设有一定的合理性,但不能涵盖所有达到目标的情况,在今后的研究中,可以将大规模用户日志与用户调查实验相结合,对用户的查询意图进行更深入的研讨。本文的下一步工作还包括对会话进行自动化划分,借助机器学习的算法和本文中人工划分得到的数据实现会话的自动划分。

参考文献:

[1] LI X, DE RIJKE M. Characterizing and predicting downloads in academic search[J]. Information processing and management, 2019, 56(3): 394－407.
[2] LI X, SCHIJVENAARS B J, DE RIJKE M, et al. Investigating queries and search failures in academic search[J]. Information processing and management, 2017, 53(3): 666－683.
[3] JIANG J, HE D, ALLAN J, et al. Searching, browsing, and clicking in a search session: changes in user behavior by task and over time[C]//International ACM SIGIR conference on research and development in information retrieval. USA: ACM, 2014: 607－616.
[4] DOGAN R I, MURRAY G C, NEVEOL A, et al. Understanding PubMed? user search behavior through log analysis[EB/OL]. [2021－03－06]. https://www.researchgate.net/publication/41435872_Understanding_PubMedR_user_search_behavior_through_log_analysis.
[5] KHABSA M, WU Z, GILES C L, et al. Towards better understanding of academic search[C]// Joint conference on digital libraries. USA:ACM,2016: 111－114.
[6] ARAMPATZIS A, KAMPS J. A study of query length[C]//International ACM SIGIR conference on research and development in information retrieval. USA:ACM,2008: 811－812.
[7] XIE K, YU H, CEN R, et al. Using log mining to analyze user behavior on search engine[J]. Frontiers of electrical and electronic engineering in China, 2011, 7(2): 254－260.
[8] KAMVAR M, KELLAR M, PATEL R, et al. Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices[C]// The Web conference. USA: ACM,2009: 801－810.
[9] LIU Y, MIAO J, ZHANG M, et al. How do users describe their information need: query recommendation based on snippet click model[J]. Expert systems with applications, 2011, 38(11): 13847－13856.
[10] WANG X, FANG Z, SUN X, et al. Usage patterns of scholarly articles on Web of Science: a study on Web of science usage count[J]. Scientometrics, 2016, 109(2): 917－926.
[11] YOO I, MOSA A S. Analysis of PubMed user sessions using a full-day PubMed query log: a comparison of experienced and nonexperienced PubMed users[EB/OL]. [2021－03－06]. https://medinform.jmir.org/2015/3/e25/.
[12] WILSON T D. Human information behavior[J]. Information science, 2000, 3(2):49－56.
[13] 李月琳,樊振佳,孙星明. 探索式搜索任务属性与信息搜索行为的关系研究[J]. 情报资料工作,2017(1):54－61.
[14] ARGUELLO J. Predicting search task difficulty[C]// European conference on information retrieval. Amsterdam: Springer, 2014: 88－99.
[15] COLE M J, HENDAHEWA C, Belkin N J, et al. User activity patterns during information search[J]. ACM transactions on information systems, 2015, 33(1): 1－39.
[16] KIM J. Task difficulty as a predictor and indicator of Web searching interaction[C]// Human factors in computing systems. USA: ACM, 2006: 959－964.
[17] MARCHIONINI G. Information-seeking strategies of novices using a full-text electronic encyclopedia[J]. Journal of the Association for Information Science and Technology, 1989, 40(1): 54－66.
[18] AULA A, KHAN R, GUAN Z, et al. How does search behavior change as search becomes more difficult[C]// Human factors in computing systems. USA:ACM, 2010: 35－44.
[19] LIU J, LIU C, COLE M, et al. Exploring and predicting search task difficulty[C]// Conference on information and knowledge management. USA: ACM, 2012:1313－1322.
[20] HASSAN A, WHITE R W, DUMAIS S T, et al. Struggling or exploring?: disambiguating long search sessions[C]// Web search and data mining. USA:ACM,2014: 53－62.
[21] ODIJK D, WHITE R W, AWADALLAH A H, et al. Struggling and success in Web search[C]//Conference on information and knowledge management. USA:ACM,2015: 1551－1560.

[22] LIU J, COLE M J, LIU C, et al. Search behaviors in different task types[C]// ACM/IEEE joint conference on digital libraries. USA: ACM,2010: 69 –78.

[23] LIU J, SARKAR S, SHAH C, et al. Identifying and predicting the states of complex search tasks[C]//Conference on human information interaction and retrieval. USA: ACM, 2020:193 –202.

[24] JANSEN B J, SPINK A, BLAKELY C, et al. Defining a session on Web search engines[J]. Journal of the Association for Information Science and Technology, 2007, 58(6): 862 –871.

[25] BEITZEL S M, JENSEN E C, CHOWDHURY A, et al. Hourly analysis of a very large topically categorized Web query log[C]// International ACM SIGIR conference on research and development in information retrieval. USA:ACM,2004: 321 –328.

[26] JANSEN B J. The methodology of search log analysis[M]// JANSEN B J, SPINK A, TAKSA I. Handbook of research on Web log analysis. Hershey, PA: Idea Group Inc. ,2008.

[27] DOWNEY D, DUMAIN S T, HORVITZ E, et al. Models of searching and browsing: languages, studies, and applications[J]. Journal of the American Society for Information Science and Technology, 2007,58(6):862 –871.

[28 HE D, GOKER A. Detecting session boundaries from Web user logs [C]// Proceedings of the BCS-IRSG 22nd annual colloquium on information retrieval research. BCS Learning & Development Ltd, 2000: 57 –66.

[29 姜婷婷,王森,高慧琴. OPAC 系统用户搜索行为日志分析——以武汉大学图书馆为例[J]. 图书情报知识,2015(5):46 –56.

[30] 张鹏翼,周翔,王军. 商品检索中的多任务识别与分析[J]. 现代图书情报技术,2016(3):1 –7.

[31] JIANG T, CHI Y, GAO H. A clickstream data analysis of Chinese academic library OPAC users' information behavior[J]. Library & information science research,2017,39(3): 213 –223.

作者贡献说明:

陈翀:提出研究问题,确定研究方案,提出研究设计,撰写论文;

王思炜:调研文献,进行数据处理与实验,撰写论文;

梁冰:修改与完善研究方案。

Research on Difficulty Measurement Method in Academic Search Based on Log Mining

Chen Chong[1] Wang Siwei[1] Liang Bing[2]

[1] School of Government Management, Beijing Normal University, Beijing 100875

[2] Institute of Scientific and Technical Information of China, Beijing 100038

Abstract: [**Purpose/significance**] Users often faced different levels of information searching difficulties in search. In order to better understand user needs and improve the retrieval system, a concise and effective method was needed to measure the difficulty of searching for information. [**Method/process**] This study took the cost of effort on time and behavior for queries as manifestation of users' information seeking difficulty. The session type was divided according to the user's behavior pattern in the session, the session type with the least cost and the query requirement was satisfied as the comparison baseline, and the cost of the baseline session was used to measure the difficulty of other session types. In order to optimize the expression model of the cost, the correlation test of the behavioral indicators of the search cost was carried out, and the behavioral characteristics with good independence and discrimination were selected by factor analysis for modeling. Using National Science and Technology Library (NSTL) logs and Sogou logs as data sets to compare the difficulty faced by users in both academic search and general search environments, as well as during the exploration process represented by and different session types. [**Result/conclusion**] In the two search systems measured in this paper, the information search difficulty faced by users is 2.30 and 1.57 respectively, and the difficulty in academic search is higher than that in general search. In the two sessions that embodied the process of academic exploration, the difficulty levels were 2.35 and 4.13 respectively. The method proposed in this paper can use simple numerical values to summarize the search difficulties with multiple influencing factors, and can be used in different types of sessions and search environments, enriching the evaluation methods of the retrieval system.

Keywords: information search difficulties search difficulty measurement search cost academic search session types